Identify Patient Sets (IPS)
De-ID version 5

Software User Manual

Document version 1.06

August 2003

Center for Biomedical Informatics

Contact Info:

Melissa Saul
412.648.9838
mis@cbmi.upmc.edu

Paul Hanbury
412.383.3062
hanbury@cbmi.upmc.edu

IPS is part of the Pittsburgh IAIMS Program at the University of Pittsburgh's Center for Biomedical Informatics. This research is supported in part by IAIMS grant 1-G08-LM06625 from the National Library of Medicine to the University of Pittsburgh.

Change Log

Edited by	Date	Version #	Details
Paul Hanbury	8/11/2003	1.06	Added documentation to guide the user in adding custom entries to De-ID's dictionary files. Also, improved the installation instructions.
Paul Hanbury	7/18/2003	1.04	Updated documentation to reflect changes in De-ID interface up to version 5.08. This includes the ability to de-identify header fields.
Paul Hanbury	6/5/2003	1.02	Updated documentation to reflect changes in De-ID interface up to version 5.04. This includes the new choices in the options menu.
Paul Hanbury	1/31/2003	1.00	Original draft of De-ID documentation for De-ID version 4.14. This is a complete rewrite, detailing the new drag-and-drop header field selector and XML input.

Contents

OVERVIEW

Why de-identify?

The Health Insurance Portability and Accountability Act of 1996, or HIPAA, has required that the use of protected health information (PHI) in research studies is not permitted without the explicit consent of the patient. However, HIPAA does allow for the creation of de-identified health information. In order for clinical researchers to use clinical data in a way that complies with HIPAA, it is necessary to de-identify the records.

Table 1 found at the end of this document lists the 18 identifiers that must be removed from clinical text in order for them to be classified as de-identified under the Safe-harbor rule. De-ID allows the user to choose among these 18 items to create documents that are compliant with the Safe-harbor rule, the limited dataset rule, or neither rule.

De-ID mechanics

De-ID uses a set of heuristics to identify the presence the 18 HIPAA identifiers within the text. Supplemental dictionaries of commonly used words, geographic locations, and popular names found in the U.S. Census are also used to locate identifiable text. The UMLS Methatheasurus is utilized to ensure that words or phrases that may be medical terms containing proper names are preserved.

De-ID replaces the identifiable text with specific tags. Names found multiple times in the report are consistently replaced with the same tag to improve readability of the report. The downside of applying De-ID is the removal of a small amount of clinical information during the de-identification process. In our work to date, we have found only minor problems.

SYSTEM REQUIREMENTS

To date, a machine with minimum requirements has not been tested, but a computer with the following specifications should be able to run De-ID.

Operating System: Microsoft® Windows® 98 or higher
Processor: 400 MHz Pentium
RAM: 64 MB
Free Hard disk space: 21 MB

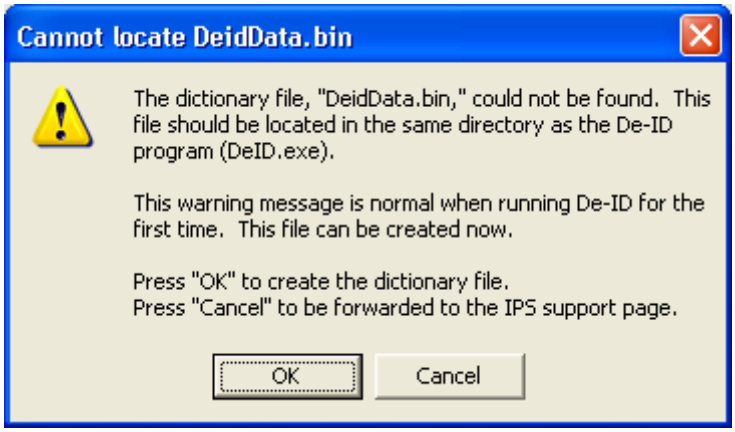
Storage of input and output files requires additional free disk space. A faster processor and more memory are recommended for better performance.

INSTALLATION

When the installation CD is placed in the drive, it should automatically start the installation CD. If this is not the case for your machine, the following steps are required.

- Double click-on “My Computer”
- Right-click the CD-ROM drive on your computer. It should contain a disk called “De-ID Install.”
- From the pop-up menu choose “Browse.”
- When the disk contents appear, double click the “Setup” program. Installation will begin.

The installation program will put several files onto your computer. This includes a De-ID executable, a LinkView executable (which is used to read linkage files), a Mix executable (which combines consecutive MARS reports belonging to the same patient), a deid_dict executable (which is used to locally update De-ID’s dictionary file), a couple of DLL files, and a directory which contains several text files (which contain all of the dictionary data used by De-ID).



The first time that De-ID is run after installation, a warning message will be displayed saying that the compressed dictionary file cannot be loaded. This is normal. Select "OK" to create this file. Once the file is created, De-ID will need to be restarted. It is impossible to run De-ID without this file; if it cannot be created, send a help request via the IPS support web site (http://support.health.pitt.edu/cs/?a=ips).

FILES

Input

There are two possible input file types. The first is the bar service header format used by MARS, Inc. at the University of Pittsburgh Medical Center (UPMC). The second is an XML format complying with our document type definition (DTD), which can be found in electronic format at the following web address: http://www.health.pitt.edu/ips/deid_in.htm. The body of this DTD is listed below.

```
<!-- Root tag, Dataset is made up of several reports -->
<!ELEMENT Dataset ( Report* ) >

<!-- Each report belongs to a patient, has a type and header (maybe) and a body -->
<!ELEMENT Report ( Patient_ID, Report_Type?, Report_Header?, Report_Text ) >

<!-- Patient_ID, should just be a study number.  Anything
  more can be added to the Report_Header and properly tagged. [ph] -->
<!ELEMENT Patient_ID ( #PCDATA ) >
<!ELEMENT Report_Type ( #PCDATA ) >


<!ELEMENT Report_Header ( Header_Person|Header_Date|Header_Data )+ >
<!ELEMENT Header_Person ( Variable, Value ) >
<!ATTLIST Header_Person role (patient|provider) "provider" >
<!ELEMENT Header_Date (Variable, Date)>
<!ELEMENT Header_Data (Variable, Value)>

<!ELEMENT Variable (#PCDATA)>
<!ELEMENT Value (#PCDATA)>


<!-- Date = Year, Month, Day with optional hours and minutes -->
<!ELEMENT Date (Year, Month, Day, Hours?, Minutes?)>
<!ELEMENT Year (#PCDATA)>
<!ELEMENT Month (#PCDATA)>
<!ELEMENT Day (#PCDATA)>
<!ELEMENT Hours (#PCDATA)>
<!ELEMENT Minutes (#PCDATA)>
<!ELEMENT Report_Text (#PCDATA)>
```

If the patients' names, health care providers' names, ID numbers or similar information is known when the documents are created, they should be supplied in the report headers. This will improve the performance of De-ID in some situations.

Output

Similarly, there are two possible output formats. The first is a plaintext format that can easily be read by the user. The second is an XML format that complies with the same DTD. This second format makes output usable in the other IPS applications.

In either output format, potential patient identifiers are replaced with tags. Each tag begins with at least two asterisks and contains a description of the data being replaced. For example, "Bob Smith" may appear as **NAME[AAA BBB]. Names found multiple times in the report are consistently replaced with the same tag to improve readability of the report. Dates are offset by some amount unknown to the user to hide the actual date but still provide a timeline of clinical events. See *Table 2* at the end of this document for a fuller overview of the tags that De-ID creates.

DE-ID INTERFACE

The Menus

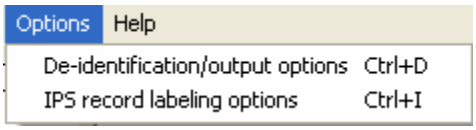
File Menu

Launch Mix: Launches the Mix application. This application can be used on MARS reports to combine multiple contiguous records into one single large record.

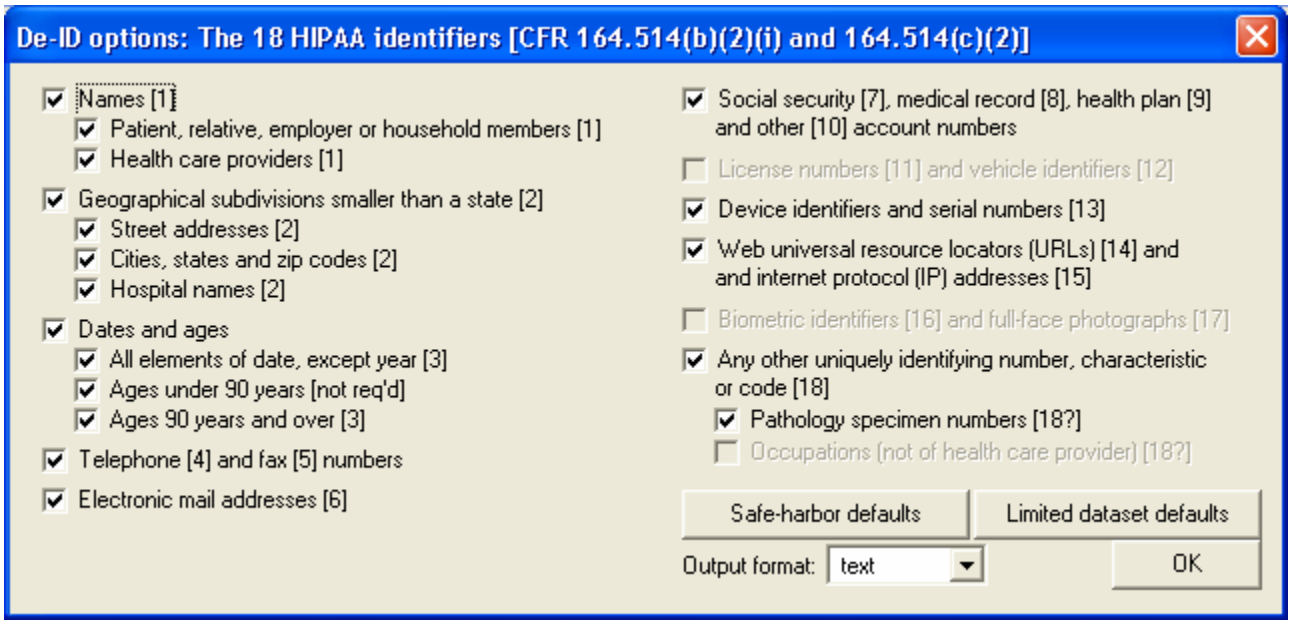
Launch Linkage File Viewer: The viewer enables the user to open a linkage file if the correct password is known.

Exit De-ID: This closes the De-ID application.

Options Menu



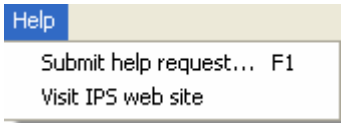
De-identification/output options: This opens a window that allows the user to select which types of personally identifying information will be stripped from the output text. This window also permits the user to select the output format of the text. This window is pictured below.



The number after each option corresponds to the Safe-harbor rules. Clicking the buttons will automatically select a configuration based upon the user's selection of either the Safe-harbor or limited dataset. The user also has the flexibility to choose a configuration that is even less strict than Limited dataset, if this is desired.

IPS record labeling options: This selection helps the user to pre-label documents for input into the IPS Retrieval Engine. Most users will not use this option.

Help Menu

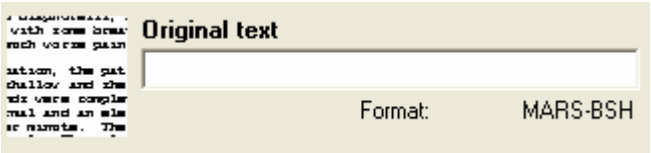


Submit help request: This opens the IPS support site (http://support.health.pitt.edu) at the Schools of the Health Sciences at the University of Pittsburgh web page. An Internet connection is required.

Visit IPS web site: This opens the IPS web site (http://www.health.pitt.edu/ips/). This site provides information about the IPS suite of applications. An Internet connection is required.

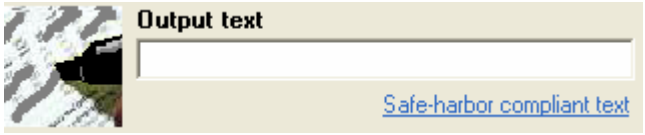
The Icons

Original text



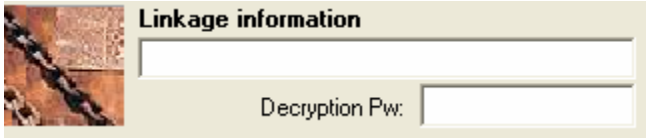
Clicking on the Original text icon to the left of the edit box will open the "File Open" dialog box. Once the user selects the input file, the program automatically detects the format. This will also select default values for Output text and Linkage information file names.

Output text



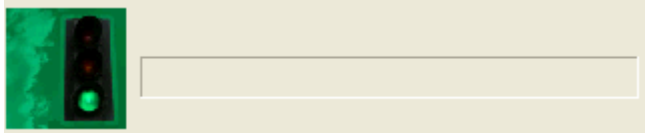
This icon is used only when the user wishes to change the default value of the output text file name. The link below the edit box indicates the level of HIPAA compliance (Safe-harbor compliant, limited dataset compliant or non-compliant) and the output format (text or XML). Clicking this link will open the output options dialog box (See Options menu section, above).

Linkage information



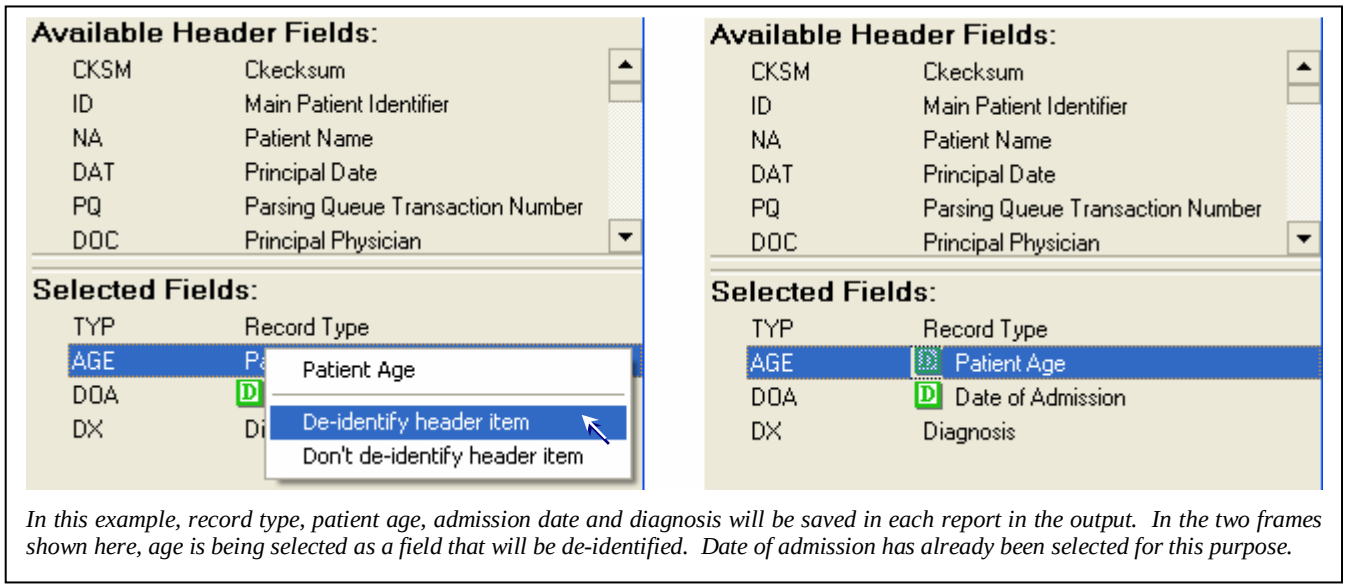
This icon is used only when the user wishes to change the default value of the linkage information file name. Linkage information will enable the user to find the patients' identity based on the de-identified headers. This file is stored in an encrypted format. The user must enter a password in the decryption password box that can be used later to decrypt this file.

Begin de-identification



Click on the green light to begin de-identification. While the file is being processed, this icon will turn into a red light, at which time the user may abort de-identification. The bar to the right of the icon indicates the progress of the de-identifier.

Header field selector



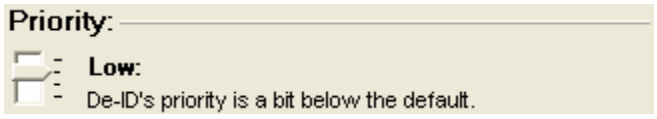
In this example, record type, patient age, admission date and diagnosis will be saved in each report in the output. In the two frames shown here, age is being selected as a field that will be de-identified. Date of admission has already been selected for this purpose.

The header field selector enables the user to choose which information should be retained from the header of the input file into the output file. Drag the field desired from the Available Header Fields and drop into Selected Fields. The selected fields will be preserved in the output file.

These header fields are not de-identified by default. If a field that contains potential identifiers is selected, this filed should be de-identified. To accomplish this, right click on the desired field and select "De-identify header item" from the popup menu. A green "D" icon will appear next to the field's name. Currently, this option is limited to name, date, age and free-text fields.

Note: For XML input, De-ID only looks at the first record to identify possible header fields. Therefore, the user should keep all potential header fields in the document, leaving blank values for those that are not used.

Priority slider



This slider allows the user to set the priority that De-ID has for system resources. The user's system will be more sluggish if the priority is set at a higher level. Possible values for this option are *Normal*, *Low*, *Lowest* and *None*. When set to *Normal*, the de-identifier will use an equal share of the processor in relation to most normal Windows applications. When set to *None*, the de-identifier will only run while all other applications are idle.

APPENDIX 1: TABLES

HIPAA Identifiers

1.	Names
2.	Geographical subdivisions smaller than a state, except the first 3 digits of zip codes; however if the region contains less than 20,000 people, the entire zip code must be replaced.
3.	All elements of dates (except year) directly relating to an individual; all ages over 89 years must be grouped into a single category of 90 or older
4.	Telephone numbers
5.	Fax numbers
6.	Electronic mail addresses
7.	Social security numbers
8.	Medical record numbers
9.	Health plan beneficiary numbers
10.	Other account numbers
11.	License numbers
12.	Vehicle identifiers
13.	Device identifiers and serial numbers
14.	Web universal resource locators (URLs)
15.	Internet protocol (IP) addresses
16.	~~Biometric identifiers~~
17.	~~Full face photographic images or any other comparable image~~
18.	Any other unique identifying number, characteristic or code

***Table 1**. According to HIPAA [CPR 164.514(2)(i)], these are the 18 identifiers that should be removed from de-identified texts. These descriptions are summarized and are not the exact HIPAA phrasing. De-ID does not make any effort in removing those items that are crossed out, as they do not normally appear in free-text reports. Currently, De-ID does not handle license numbers and vehicle identifiers.*

De-ID Tags

1.	****NAME** The name tag includes a set of letters indicating the name that has been replaced. For example, one name may be replaced with the tag, "**NAME[AAA]" and another name replaced with "**NAME[BBB CCC]." If the first name is repeated later in the document, it will be replaced again with "**NAME[AAA]." Caveat: If the same name repeats, even when belonging to a different person, the same letters will be used to denote that name. For example, Robert Johnson and Robert Williams may be replaced with "**NAME[AAA BBB]" and "**NAME[AAA CCC]," respectively.
2.	****PLACE** Names of US cities and towns are replaced with the tag **PLACE. ****INSTITUTION** Names of businesses and/or buildings (such as "Children's Hospital") are replaced with this tag. ****STREET-ADDRESS** Street numbers and names are replaced with this tag. ****ZIP-CODE** The full zip code is replaced no matter how large the population of the area happens to be.
3.	****DATE** Dates are offset by some unspecified amount, within plus or minus one year of the actual date. The date offset differs between patients based on the patients' ID numbers. Since ID numbers must be removed from de-identified reports, it is not possible for the end user to determine the offset based on this information. Also, since ID numbers do not (usually) change for a patient during her treatment, the timeline of events within this treatment should remain consistent. ****AGE** De-ID rounds ages to the nearest decade. Thus a 35-year-old will be given the tag **AGE[in 30s]. There are three exceptions to this rule. Two of these are for children where the ranges "**AGE[birth-12]" and "**AGE[in teens]" are used; and one is for the elderly where "**AGE[90+]" is used.
4.	****PHONE** Telephone numbers are replaced with the tag **PHONE.
5.	****PHONE** Fax numbers are not distinguished from telephone numbers, and use the same tag.
6.	****EMAIL** De-ID finds email addresses in any of the following domains: com, net, gov and edu.
7.	****ID-NUM**
8.	****ID-NUM**
9.	****ID-NUM**
10.	****ID-NUM**
11.	Due to a lack of training data in the MARS database, De-ID does not currently handle license numbers
12.	Due to a lack of training data in the MARS database, De-ID does not currently handle vehicle identifiers.
13.	****DEVICE-ID** De-ID removes serial and model numbers, but not model names.

14.	****WEB-LOC** De-ID finds Internet URLs in any of the following domains: com, net, gov and edu.
15.	****WEB-LOC**
16.	Biometric identifiers are not a part of free-text reports and are not handled by De-ID.
17.	Images are not a part of free-text reports and are not handled by De-ID.
18.	****PATH-NUMBER** Pathology specimen numbers are replaced with this tag.

***Table 2**. The numbers in this table correspond with the numbers in* Table 1.

APPENDIX 2: EDITING THE DICTIONARY

From De-ID's installation directory (C:\Program Files\U-Pitt\IPS\, by default) there is a subdirectory named data. This subdirectory contains several files that are incorporated into De-ID's dictionary.

corp_suffixes.txt	These are words that give clues that the name of a hospital and business occurs in the preceding tokens. (e.g., Hospital or Inc.)	W
cuw.txt	These are commonly used words in clinical reports. In most cases De-ID will not scrub these words when an ambiguity occurs.	P
drugs.txt	These are drug names. De-ID will not usually scrub any phrases on this list.	P
hc_suffixes.txt	These are name suffixes for health care employees. (e.g., M.D. or R.N.)	W
hc_titles.txt	These are titles for health care employees. (e.g., Dr. or Nurse)	W
months.txt	These are the months of the year with abbreviated versions. (e.g., July or Jul.)	W
names.txt	These are the most common names in the US according to the 2000 census. This list includes both last and first names.	W
places.txt	These are the names of the most populated US towns according to the 2000 census.	P
pt_suffixes.txt	These are name suffixes that are not necessarily health care related. (e.g. Jr. or Sr.)	W
pt_titles	These are non-health care related name titles. (e.g., Gen. or Rev.)	W
states.txt	This is a list of US states with postal abbreviations. (e.g., Ohio or OH)	P
stop_words.txt	This is a standard list of stopwords.	W
street_suffixes.txt	This are words that give clues that the name of street occurs in the preceding tokens. (e.g., Ave. or Road.)	W
umls.txt	These are clinically meaningful terms according to the UMLS Metathesaurus. These phrases, like the drug name phrases, are rarely de-identified.	P

***Table 3**. The first column contains the name of each file, the second column contains a description of what the file contains, and the third column tells whether the entries in the file are words (W) or phrases (P). Most of these files will never need to be changed by the user. There are a few problems, however, that can be fixed by changing one of these files.*

Dictionary file format

Each of these files is a plain ASCII text file containing one entry per line. Typical UNIX-style comments are recognized. That is, any "#" character begins a comment and continues to the end of the line. Care should

be taken to not include multi-word entries in files that should only contain single words, as the program will not issue a warning if this happens.

```
# This is a comment starting at the beginning of the line.
Word                 # This is a single word entry
Three word phrase    # This is a phrase entry
# The two preceding lines contain midline comments.
# The entries “Word” and “Three word phrase” will be added to De-ID’s
# dictionary file.  None of the comments are added to the dictionary file.
```

These text files take a long time to load into memory, so, to allow De-ID a faster startup time, must be compressed into a binary format. To do this, simply run the program deid_dict.exe, located in De-ID’s installation directory (C:\Program Files\U-Pitt\IPS\, by default). This will recreate the deiddata.bin file.

We recommend making such changes at the end of the data file and annotating them, as the current version of De-ID does not “remember” these changes when upgrading.

Making changes

Preventing overmarks

Sometimes a clinically meaningful term is misidentified as being a name. In order to prevent this from happening, it could be added to one of the cuw.txt or umls.txt files.

Preventing undermarks

If a name is not scrubbed, it name may be added to the names.txt file. This will increase the likelihood that De-ID will properly recognize that word as a name in the future.

If a location is not scrubbed, that location can be added to the places.txt file. This will increase the likelihood that De-ID will properly recognize the location name. This works equally as well for corporate names, but will cause the tag **PLACE to be used instead of **INSTITUTION.

Correcting omissions

There may be different words that are not included by default in one of the files that should be. This may include adding words to a file or recreating an entire file based on your own needs.

The file cuw.txt was automatically generated using reports from the Presbyterian University Hospital of the University of Pittsburgh Medical Center (UPMC). This file suits the needs for researchers at UPMC, but may need to be updated to meet other institutions’ needs.